ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

SENTIMENT ANALYSIS OF 2019 PRESIDENTIAL CANDIDATES ON KOMPAS.COM THROUGH TOPIC-DRIVEN CRAWLERS

Muhammad Nasar
Teknik Elektro/Universitas Muhammadiyah Malang

Kontak Person:
Muhammad Nasar
Jl. Raya Tlogomas 246 Malang, Telp/Fax 0341-464318 ext 129
E-mail: nasar@umm.ac.id

Abstrak

Web merekam sejumlah besar data yang berpotensi digali untuk mengungkap informasi yang masih tersembunyi, misalnya analisis sentimen. Namun web memiliki ukuran yang sangat besar, tidak terstruktur, dan terus berkembang. Mengingat besarnya volume data web, membangun data warehouse untuk keperluan analisis sentimen bukanlah solusi efisien. Karena analisis sentimen hanya membutuhkan topik tertentu untuk digali. Penelitian ini bertujuan (1) mendesain crawler guna mengambil halaman web topik tertentu untuk analisis sentimen, (2) mendesain filter text preprocessing untuk membersihkan teks target dari kode HTML, dan (3) menganalisis sentimen berita terkait Capres-Cawapres 2019 pada situs kompas.com. Peneliti mencoba mengonstruksi crawler sederhana memanfaatkan mesin pencari Google. Crawler ini bekerja 2 tahap: mengambil URL indeks, dan mengambil halaman target. Cara ini berhasil mengumpulkan daftar URL sesuai topiknya. Output crawler berupa teks bercampur kode HTML di filter sedemikian rupa sehingga teks target bisa dibersihkan. Teks yang sudah dibersihkan kemudian diproses menggunakan Naïve Bayes Classifier untuk mengetahui polaritas sentimennya. Hasil percobaan menunjukkan teknik ini cukup efektif. Crawler hanya mengambil halaman web sesuai topik, filter dapat membersihkan teks target, dan polaritas sentimen dapat diketahui. Dari 17 berita di kompas.com yang dianalisis, kedua pasangan Capres-Cawapres telah diberitakan positif, namun Jokowi-Ma'ruf cenderung diberitakan lebih positif (85,5%) daripada Prabowo-Sandi (46,8%).

***Kata kunci**: Web Crawler, Web Mining, Big Data Analytics, Sentiment Analysis*

Abstract

Web records a large amount of data potential to be mined to reveal hidden information, e.g. sentiment analysis. However, since the web has a very large, unstructured, and growing data, building a data warehouse for sentiment analysis is not an efficient choice. This is because sentiment analysis only requires certain topics to be explored. This study aims to (1) design a crawler for taking certain topics on web pages for sentiment analysis, (2) design text preprocessing filter to clean the targeted text from HTML code, and (3) measure the sentiment of 2019 presidential candidates on the Kompas.com site. I tried to construct a simple crawler using Google search engine. This crawler works in two stages: picking up the indexed URL, then picking up the targeted page. This method successfully collects a list of URLs according to a given topic. The crawler's output is text mixed with HTML code filtered so that the targeted text can be cleaned. The clean text is then processed using the Naïve Bayes Classifier to determine the polarity of its sentiment. The experimental results show this technique is quite effective. Crawler only takes web pages according to the given topics, filter is able to clean the targeted text, and sentiment polarity can be measured. Of 17 news crawled from the Kompas.com, both presidential candidates have been reported positively, nevertheless Jokowi-Ma'ruf tends to be more positive than Prabowo-Sandi, 85.5% and 46.8% respectively.

***Keywords**: web crawler, web mining, big data analytics, sentiment analysis*

Pendahuluan

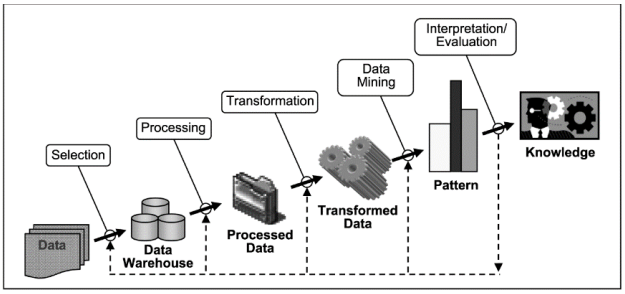
Era industri 4.0 ditandai dengan masifnya penggunaan media informasi digital. Saat ini transformasi dunia digital (online) sudah berada pada periode *social web*, kondisi dimana web populer digunakan untuk mendukung dan memupuk interaksi sosial. Jutaan orang terhubung satu sama lain, berkomunikasi, dan berbagi informasi. Mereka membaca, berdiskusi, menulis opini, menilai, dan mengekspresikan pendapatnya melalui web [1],[2]. Interaksi yang terjadi pada platform internet ini akan terus meningkat dan menghasilkan maha data *(big data)* yang berpotensi untuk digali. Pola-pola tersembunyi yang tadinya sulit ditemukan dengan kasat mata, bisa diketahui dan divisualkan. Di Indonesia sendiri, pertumbuhan web didominasi perusahaan *startups*, *e-commerce*, dan berbagai media pemberitaan online. Bahkan perusahaan pemberitaan yang sebelumnya berbasis kertas kini sudah online. Salah satunya adalah kompas.com.

Penggunaan web yang terus meningkat termasuk jurnalistik online ini menarik untuk diteliti. Salah satunya melihat sentimen berita di dalamnya. *Sentiment analysis* adalah bidang studi data mining

ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

untuk menganalisis opini penulisnya terhadap suatu entitas seperti produk, layanan, organisasi, individu, atau topik tertentu. Ada beberapa variasi istilah yang sedikit berbeda, misalnya analisis sentimen, penambangan opini, ekstraksi pendapat, analisis subjektivitas, *review mining*, dll. Namun *sentiment analysis* sudah umum digunakan [3]. Dengan analisis sentimen, suatu opini dalam bahasa tulisan dapat dievaluasi apakah ekspresi penulisnya positif, netral, atau negatif [3]. Data mining sendiri merupakan disiplin dalam pengumpulan, penyimpanan, dan penggalian data untuk menemukan pola berharga yang tersembunyi. Data mining umumnya melibatkan *data warehouse* untuk menyimpan data [4]. Sumber data mining dapat berasal dari sensor, operator, mesin, web, dan sebagainya [5]. Penambangan data melalui web disebut juga *web mining*, yang menjadi fokus dalam penelitian ini. *Web mining* adalah salah satu teknik data mining yang mengekstrak informasi dari web server. Sumber data ini meliputi halaman web, link, objek, dan log web server [6],[7].

Namun, melakukan analisis sentimen berita resmi melalui *web mining* akan mengalami beberapa kendala jika dilakukan dengan teknik data mining biasa. Pertama, web memiliki ukuran data yang sangat besar dan terus berkembang. Atas pertimbangan ini, membangun *data warehouse* normalnya dalam *data mining* untuk penyimpanan data web secara random bukanlah solusi efisien. Karena analisis sentimen hanya membutuhkan topik tertentu. Tahapan data mining ditunjukkan pada Gambar 1.



Gambar 1 Tahapan data mining [4]

Kedua, situs berita umumnya tidak menyediakan *application programming interface (API)* untuk akses data mesin-ke-mesin. Media social Twitter dan Facebook menyediakan *API* untuk mengakses data terstruktur ke servernya, namun situs pemberitaan seperti kompas.com tidak menyediakan fitur ini. Kompas.com dan kebanyakan situs berita lainnya hanya menyediakan halaman yang bersifat searah, di optimasi untuk dibaca manusia. Konsekuensinya, pelaku *web mining* perlu mengembangkan teknik tersendiri agar data bisa diambil. Program untuk mengambil data dari halaman web ini biasa disebut dengan *crawler* [8]. Sebuah *web crawler* dapat secara sistematis menjelajahi suatu website, mengambil, dan menyimpannya ke database lokal. *Topic-driver crawler* adalah konsep pendekatan pengambilan halaman web yang mengandung topik tertentu saja [9].

Permasalahan ketiga terkait struktur *uniform resource locator* (URL). Crawler bekerja dengan terlebih dahulu mengidentifikasi URL target. URL berisi informasi alamat server di internet berikut susunan direktori dan nama file. Sebuah URL a.b.c/d/e.html dapat diartikan komputer klien akan mengakses server a pada domain b.c dengan target file e.html pada direktori /d. Namun sejak URL dibentuk berdasarkan nama yang unik, maka pola struktur setiap website akan berubah tergantung gaya pengelolanya. Sedangkan tantangan keempat, suatu dokumen web tentu mengandung banyak sekali kode *HTML (hyper text markup language)* dan *script* lainnya. Kode tersebut digunakan untuk mempertahankan format tampilan agar nyaman dibaca manusia. Struktur kode HTML ini juga tidak tetap tiap halamannya. Teks target dan teks lain seperti iklan, link gambar, bercampur menjadi satu halaman file. Agar proses data mining berjalan efektif, segala karakter selain teks target perlu dibersihkan terlebih dahulu.

Beberapa studi terkait analisis sentimen berbasis *web mining* sudah banyak dilakukan, baik terhadap produk elektronik, film, travel, dan lainnya. Bing Lu, dkk. [10] menganalisis perbandingan review berbagai produk elektronik pada website dengan membangun sebuah prototipe dinamakan *opinion observer*. Sistem ini mengambil review pelanggan (pro dan kontra konsumen) dari halaman web *(web crawling)* sebagai sumber data. Data training dibentuk dengan melakukan tagging secara manual pada sejumlah review atas 15 produk elektronik. Pranali Yenkar [11] menganalisis review penonton film

ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

melalui web blog. Senada dengan peneliti sebelumnya, peneliti ini juga memanfaatkan crawler untuk mengambil data di web. Namun crawler yang digunakan adalah aplikasi open source yang sudah jadi, yaitu OpenWebSpider dan Arachnode. Crawler ini mengambil semua isi website tanpa membedakan topik yang diperlukan saja sehingga dianggap masih kurang efisien.

Crawler sendiri pernah diteliti oleh [9],[12] yang mengklasifikasikan crawler menjadi *traditional crawler, deep crawler*, dan *RIA (Rich Internet Application). Traditional crawler* memerlukan input berupa URL dan akan menelusuri setiap *hyperlink* yang ada di dalamnya. Tipe ini mengasumsikan semua konten dari aplikasi web dapat dijangkau melalui URL. *Deep crawler* diperuntukkan untuk halaman web yang interaktif, dimana URL memerlukan input form untuk operasinya. RIA merupakan tipe *crawler* yang dapat berjalan pada mode *client-side*. Sedangkan [8] mengembangkan *parallel crawler* untuk meningkatkan kapasitas dan skalabilitas, sebagaimana ukuran web yang saat ini tumbuh makin besar. Sedangkan sebagian besar studi lainnya tidak memerlukan *crawler* karena menggunakan API untuk *data harvesting*.

Dari uraian di atas dapat diketahui bahwa beberapa penelitian terdahulu sudah melakukan analisis sentiment suatu topik menggunakan data yang bersumber dari web. Namun penelitiannya (1) hanya fokus pada metode klasifikasi, dan (2) jenis sumber data adalah opini perorangan (bukan berita resmi yang ditulis jurnalis). Di samping itu, (3) tidak dibahas secara spesifik mengenai proses crawler yang digunakan, serta tantangan-tantangan dalam pembersihan data web. Studi tentang crawler sendiri cenderung diarahkan untuk penggalian data web skala besar, bukan spesifik untuk *sentiment analysis*.

Untuk itulah dalam penelitian ini pertama penulis fokus pada desain crawler yang efisien dalam melakukan pengambilan halaman web, yaitu dengan mengambil topik tertentu saja. Kedua, penulis juga menunjukkan desain crawler dan desain text preprocessing untuk membersihkan teks HTML hasil *crawling*. Dan ketiga, jenis data yang diproses untuk dianalisis dalam penelitian ini berupa berita resmi, yaitu berita yang diambil langsung dari situs kompas.com. Desain *crawler* dan *text preprocessing* ini akan dibahas pada bagian metodologi.

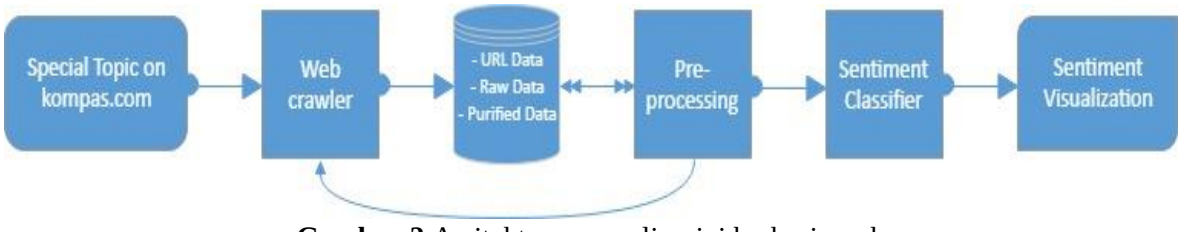
Pemilihan sumber data dan topik juga memiliki alasan tersendiri. Situs kompas.com dipilih karena pertimbangan kelengkapan topik berita, kemudahan identifikasi pola, dan popularitas situs. Sedangkan topik Capres-Cawapres (kependekan dari calon presiden dan calon wakil presiden) 2019 dipilih mengingat saat penelitian ini dilakukan, Indonesia tengah memasuki tahun politik, dimana isu ini sedang hangat diperbincangkan dan tentunya menjadi perhatian khusus bagi banyak pihak. Tidak lama lagi Indonesia akan menyelenggarakan pesta demokrasi pemilihan presiden dan wakil presiden periode 2019-2024.

Dengan demikian, penelitian ini diharapkan dapat berkontribusi di bidang *web mining*, khususnya menawarkan model alternatif *crawler* dan *text preprocessing* yang efisien untuk keperluan analisis sentimen berbasis web. Selain itu, dengan topik Capres-Cawapres 2019 diharapkan dapat memberi gambaran kepada masyarakat luas bagaimana teknologi *web mining* dimanfaatkan untuk menggali informasi politik secara saintifik.

2. Metode Penelitian

Penelitian ini didesain untuk mengolah berita formal pada harian kompas.com, dan menganalisis apakah sentimennya cenderung positif, netral, atau negatif. Data yang diolah adalah teks berbahasa Indonesia bertopik Capres-Cawapres 2019. Kata kunci "jokowi maruf" digunakan untuk pasangan Joko Widodo dan Ma'ruf Amin, dan "prabowo sandi" untuk pasangan Prabowo Subianto dan Sandiaga Uno.

Ada 3 tahapan proses dalam sistem ini, yaitu (1) *crawling* dan penyimpanan data, (2) pembersihan data, dan (3) klasifikasi sentimen. Tahapan proses ini didesain sebagaimana Gambar 2. Desain algoritma *crawling*, *text preprocessing*, dan *sentiment classifier* dijelaskan setelahnya.



Gambar 2 Arsitektur penggali opini berbasis web

ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

2.1 Algoritma Crawling dan Pre Processing

Proses *crawling* halaman berita bertopik baik Joko Widodo – Ma'ruf Amin maupun Prabowo Subianto – Sandiaga Uno pada situs kompas.com adalah sebagai berikut:

1. Web crawler melakukan request HTTPS GET melalui mesin pencari Google standar dengan format query sebagai berikut: <kata+kunci> site:kompas.com
2. Web crawler akan menerima respon halaman pertama indeks berisi tidak lebih dari 10 URL target berita relevan. Indeks disimpan di database URL Data. URL iklan tidak diambil.
3. Indeks URL berupa file HTML dibaca oleh blok *preprocessing* untuk dibersihkan dan diambil daftar URL beritanya saja. Identifikasi pola halaman indeks hasil pencarian google ditunjukkan pada Gambar 3.

Gambar 3 Identifikasi tag URL halaman hasil pencarian google

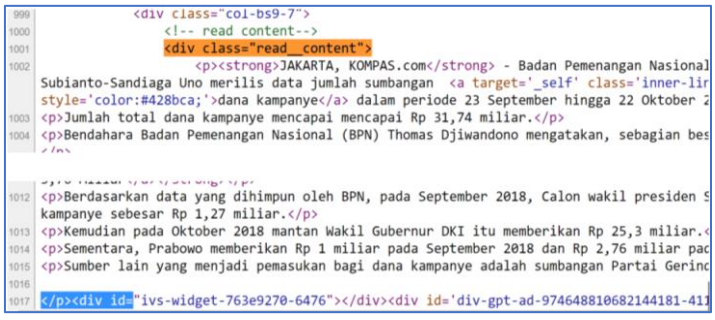
Dari pola halaman hasil pencarian google pada Gambar 3 dapat diidentifikasi keyword URL target adalah berada setelah tag . Maka rancangan algoritma pembersihan teks halaman web ini dalam bentuk *pseudocode* ditunjukkan pada Gambar 4.

```
Kenali keyword link target
Hapus seluruh karakter sebelum link target pertama
Hitung jumlah link target
While jumlah link target lebih dari 0
    Jika jumlah link target lebih dari 1
      Bersihkan seluruh karakter antar link target
    Jika jumlah link target sama dengan 1
      Bersihkan seluruh karakter setelah link target terakhir
Kurangi nilai jumlah link target dengan 1
End while
```

Gambar 4 Desain algoritma pembersihan link target pada hasil pencarian google

Contoh URL hasil ekstrak halaman indeks hasil *crawling* ditunjukkan pada Tabel 3.

1. *Crawler* akan menggunakan daftar URL target untuk mengambil isi beritanya, lalu disimpan sebagai raw data.
2. Teks berita pada raw data (masih bercampur kode HTML) dibersihkan oleh blok *pre processing*. Pola kode halaman berita kompas.com ditunjukkan pada Gambar 5.
3. Teks berita yang sudah dimurnikan disimpan di database sebagai *purified data* dan siap untuk dianalisis.



Gambar 5 Identifikasi tag berita halaman Kompas.com

Dari pola kode halaman berita kompas.com pada Gambar 5, dapat diidentifikasi bawa teks target berada setelah tag dan sebelum tag </p><div id=. Dari pola ini dirancang algoritma *text cleansing* seperti Gambar 6.

ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

```
Kenali keyword awal teks target
Kenali keyword akhir teks target
Hapus seluruh karakter sebelum teks target
Hapus seluruh karakter sesudah teks target
Hapus seluruh kode HTML pada teks target
End
```

Gambar 6 Desain algoritma pembersihan teks halaman berita Kompas.com

2.2 Proses Analisis Sentimen

Dalam proses klasifikasi, sentiment dikelompokkan ke dalam 3 kategori, positif, netral, dan negatif. Pro dan kontra, suka dan tidak suka, kata atau frasa yang baik dan buruk terlebih dahulu dikategorikan dalam sebuah basis data, disebut juga *dictionary. Dictionary* ini kemudian difungsikan sebagai data latih. Metode Naïve Bayes bekerja dengan memprediksi probabilitas keputusan berdasarkan pengalaman sebelumnya pada data latih. Probabilitas tertinggi dianggap sebagai keputusan yang paling mungkin. Rumusan Naïve Bayes ditunjukkan pada Persamaan 1.

$$P(X|Y) = (P(X|Y) \times P(X))/P(Y) \quad (1)$$

Persamaan tersebut menunjukkan bahwa probabilitas kejadian X sebagai Y ditentukan dari peluang Y saat X dikalikan dengan peluang X, dan dibagi dengan total kejadian Y. Dalam *machine learning*, Naïve Bayer tergolong *supervised learning* karena membutuhkan data training dalam keputusannya [5],[13],[14]. Sebagai ilustrasi, anggaplah ada 10 baris data training sebagaimana Tabel 1.

Tabel 1 Contoh data training [15]

ID	X1	X2	X3	Y
1	Yes	Single	125	No
2	No	Married	100	No
3	No	Single	70	No
4	Yes	Married	120	No
5	No	Divorce	95	Yes
6	No	Married	60	No
7	Yes	Divorce	220	No
8	No	Single	85	Yes
9	No	Married	75	No
10	No	Single	90	Yes

Dari data training tersebut dapat dihitung Y untuk X1=no, X2=married, dan X3=120, sebagaimana data testing pada Tabel 2.

Tabel 2 Contoh data testing

ID	X1	X2	X3	Y
1	No	Married	120	?

Berdasarkan Persamaan 1 dan data training pada Tabel 1, maka probabilitas Y pada data testing ini dapat diselesaikan dengan langkah-langkah berikut:

1. Nilai probabilitas $P(Y = yes) = \frac{3}{10}$ dan $P(Y = no) = \frac{7}{10}$
2. Nilai probabilitas $P(X1 = no|Y = yes) = \frac{3}{3}$ dan $P(X1 = no|Y = no) = \frac{4}{7}$
3. Nilai probabilitas $P(X2 = married|Y = yes) = \frac{0}{3}$ dan $P(X2 = married|Y = no) = \frac{4}{7}$
4. Nilai probabilitas $P(X3 = 120|Y = yes) = \frac{0}{3}$ dan $P(X3 = 120|Y = no) = \frac{1}{7}$

ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

Setelah mengenali nilai masing-masing probabilitas, maka Persaman 1 menunjukkan kemungkinan $Y = yes$.

$$P(yes) = P(X1 = no|Y = yes) \times P(X2 = married|Y = yes) \times P(Y = yes) \\ = \frac{3}{3} \times \frac{0}{3} \times \frac{0}{3} \times \frac{3}{10} = 0 \quad (2)$$

Pada Persamaan 3 kemungkinan $Y = no$.

$$P(no) = P(X1 = no|Y = no) \times P(X2 = married|Y = no) \times P(Y = no) \\ = \frac{4}{7} \times \frac{4}{7} \times \frac{1}{7} \times \frac{7}{10} = 0.033 \quad (3)$$

Karena $P = no$ lebih besar dari $P = yes$, maka untuk $X(1 = No, X2 = married, X3 = 120$, keputusan Y adalah "tidak", atau $Y = no$.

Prinsip yang sama digunakan untuk menghitung kelompok kata pada sebuah kalimat, paragraf, atau dokumen berita pada penelitian ini. Daftar kata-kata positif, netral, dan negatif terlebih dahulu dikelompokkan kedalam basis data atau *dictionary* sebagai data training. *Classifier* pada penelitian ini menggunakan PhpInsight berbahasa Indonesia, termasuk data training yang ada di dalamnya. PhpInsight merupakan *software* pengklasifikasi sentimen berbasis Naïve Bayes ditulis pertama kali oleh James Hennessey (2012) dan disebar bebas di internet [16]. Dalam ini, fitur visualisasi grafik pie chart ditambahkan agar data lebih mudah dibaca.

2.3 Metode Pengujian

Pengujian dilakukan menggunakan CPU Intel Core i7, memory 16Gb, sistem operasi Windows 10 64bit, webserver Apache 2.4.25, PHP 5.6.30, dan database MySQL 5.0, terhubung ke internet berkisar 2Mbps simetrik. Prosedur pengujian dilakukan dengan mengambil indeks URL halaman pertama pencarian google untuk masing-masing pasangan calon. Dalam hal ini *crawler* akan dipakai bergantian. Tidak ada pengaruh apakah kata kunci "jokowi maruf" atau "prabowo sandi" yang didahulukan.

3. Hasil Penelitian dan Pembahasan

Pada saat percobaan ini, *crawling* kata kunci "jokowi maruf" menghasilkan 8 indeks URL dan kata kunci "prabowo sandi" menghasilkan 9 indeks URL, masing-masing ditunjukkan pada Tabel 3 dan 4.

Tabel 3. Hasil crawling indeks URL terkait "jokowi maruf"

ID URL	URL
1	https://regional.kompas.com/read/2018/11/11/09242121/dedi-mulyadi-jokowi-maruf-amin-pasti-menang-telak-di-jawa-barat
2	https://nasional.kompas.com/read/2018/11/05/19191251/yusril-jadi-pengacara-jokowi-maruf
3	https://nasional.kompas.com/read/2018/11/11/07453131/jokowi-maruf-targetkan-kemenangan-60-persen-di-jawa-barat
4	https://regional.kompas.com/read/2018/11/11/12123241/maruf-setiap-hari-ada-deklarasi-untuk-jokowi-amin-di-jawa-barat
5	https://regional.kompas.com/read/2018/11/11/20375631/surya-paloh-yakin-jokowi-maruf-menang-salah-satu-lumbungnya-jateng
6	https://nasional.kompas.com/read/2018/11/04/18465991/rumah-aspirasi-jokowi-maruf-dibuka-untuk-umum-apa-saja-kegiatannya
7	https://nasional.kompas.com/read/2018/09/28/12414821/atribut-kampanye-jokowi-maruf-amin-diluncurkan-pada-oktober-2018

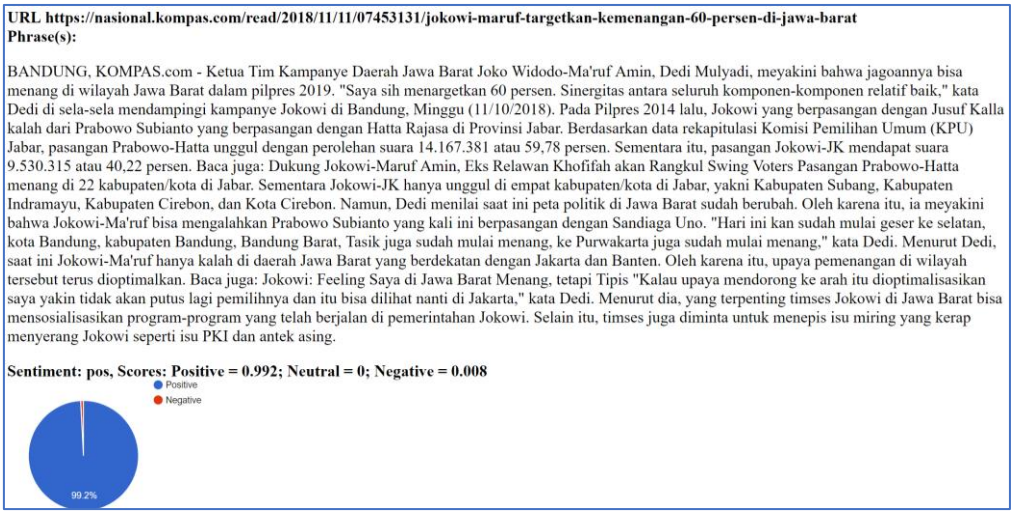
ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

8	https://nasional.kompas.com/read/2018/10/24/06552501/survei-kompas-jokowi-maruf-526-persen-prabowo-sandi-327-persen

Tabel 4. Hasil crawling indeks URL terkait "prabowo sandi"

ID URL	URL
1	https://nasional.kompas.com/read/2018/11/09/21184081/anak-anak-orasi-lupakan-jokowi-di-aksi-211-prabowo-sandi-akan-dilaporkan-ke
2	https://nasional.kompas.com/read/2018/10/24/06552501/survei-kompas-jokowi-maruf-526-persen-prabowo-sandi-327-persen
3	https://nasional.kompas.com/read/2018/10/23/19393831/oktober-ini-prabowo-sandi-dapat-dana-kampanye-rp-12-juta-dari-pendukung
4	https://nasional.kompas.com/read/2018/11/02/11142771/prabowo-sandi-pilih-cucu-pendiri-nu-jadi-jubir-peringatan-untuk-jokowi
5	https://nasional.kompas.com/read/2018/11/02/14500381/kubu-jokowi-sebut-jubir-baru-prabowo-sandi-tak-signifikan-gaet-suara-nu
6	https://nasional.kompas.com/read/2018/10/24/17121321/baru-gerindra-yang-sumbang-dana-untuk-prabowo-sandi-ini-kata-demokrat
7	https://nasional.kompas.com/read/2018/10/23/20163681/hingga-kini-hanya-gerindra-yang-sumbang-dana-kampanye-prabowo-sandi
8	https://megapolitan.kompas.com/read/2018/10/26/19494201/hoaks-undangan-cfd-membiru-prabowo-sandi-yang-disebar-di-medsos
9	https://nasional.kompas.com/read/2018/11/05/09042771/ketika-prabowo-merasa-candaannya-selalu-dipermasalahkan

Tabel 3 dan 4 di atas menunjukkan bahwa *crawler* berhasil mengambil URL yang relevan dengan topik yang diinginkan. Namun Tabel 3 berisi 8 URL sedangkan Tabel 4 berisi 9. Perbedaan ini normal terjadi mengingat pencarian URL dilakukan didasarkan topik yang paling relevan saja. Filter *preprocessing* juga berhasil membersihkan karakter lain selain URL target. Sebagaimana tahapan yang sudah dijelaskan sebelumnya, setelah indeks URL dibersihkan, dilakukan pengambilan halaman web berdasarkan indeks URL di atas. Pada tahap ini, halaman web berita kompas.com juga berhasil diambil dan dibersihkan sehingga teks berita dapat dianalisis. Gambar 7 dan 8 menunjukkan bahwa sentimen berita yang terdiri dari ratusan kata tersebut berhasil dianalisis.


URL https://nasional.kompas.com/read/2018/11/11/07453131/jokowi-maruf-targetkan-kemenangan-60-persen-di-jawa-barat
Phrase(s):

BANDUNG, KOMPAS.com - Ketua Tim Kampanye Daerah Jawa Barat Joko Widodo-Ma'ruf Amin, Dedi Mulyadi, meyakini bahwa jagoannya bisa menang di wilayah Jawa Barat dalam pilpres 2019. "Saya sih menargetkan 60 persen. Sinergitas antara seluruh komponen-komponen relatif baik," kata Dedi di sela-sela mendampingi kampanye Jokowi di Bandung, Minggu (11/10/2018). Pada Pilpres 2014 lalu, Jokowi yang berpasangan dengan Jusuf Kalla kalah dari Prabowo Subianto yang berpasangan dengan Hatta Rajasa di Provinsi Jabar. Berdasarkan data rekapitulasi Komisi Pemilihan Umum (KPU) Jabar, pasangan Prabowo-Hatta unggul dengan perolehan suara 14.167.381 atau 59,78 persen. Sementara itu, pasangan Jokowi-JK mendapat suara 9.530.315 atau 40,22 persen. Baca juga: Dukung Jokowi-Maruf Amin, Eks Relawan Khofifah akan Rangkul Swing Voters Pasangan Prabowo-Hatta menang di 22 kabupaten/kota di Jabar. Sementara Jokowi-JK hanya unggul di empat kabupaten/kota di Jabar, yakni Kabupaten Subang, Kabupaten Indramayu, Kabupaten Cirebon, dan Kota Cirebon. Namun, Dedi menilai saat ini peta politik di Jawa Barat sudah berubah. Oleh karena itu, ia meyakini bahwa Jokowi-Ma'ruf bisa mengalahkan Prabowo Subianto yang kali ini berpasangan dengan Sandiaga Uno. "Hari ini kan sudah mulai geser ke selatan, kota Bandung, kabupaten Bandung, Bandung Barat, Tasik juga sudah mulai menang, ke Purwakarta juga sudah mulai menang," kata Dedi. Menurut Dedi, saat ini Jokowi-Ma'ruf hanya kalah di daerah Jawa Barat yang berdekatan dengan Jakarta dan Banten. Oleh karena itu, upaya pemenangan di wilayah tersebut terus dioptimalkan. Baca juga: Jokowi: Feeling Saya di Jawa Barat Menang, tetapi Tipis "Kalau upaya mendorong ke arah itu dioptimalisasikan saya yakin tidak akan putus lagi pemilihnya dan itu bisa dilihat nanti di Jakarta," kata Dedi. Menurut dia, yang terpenting timses Jokowi di Jawa Barat bisa mensosialisasikan program-program yang telah berjalan di pemerintahan Jokowi. Selain itu, timses juga diminta untuk menepis isu miring yang kerap menyerang Jokowi seperti isu PKI dan antek asing.

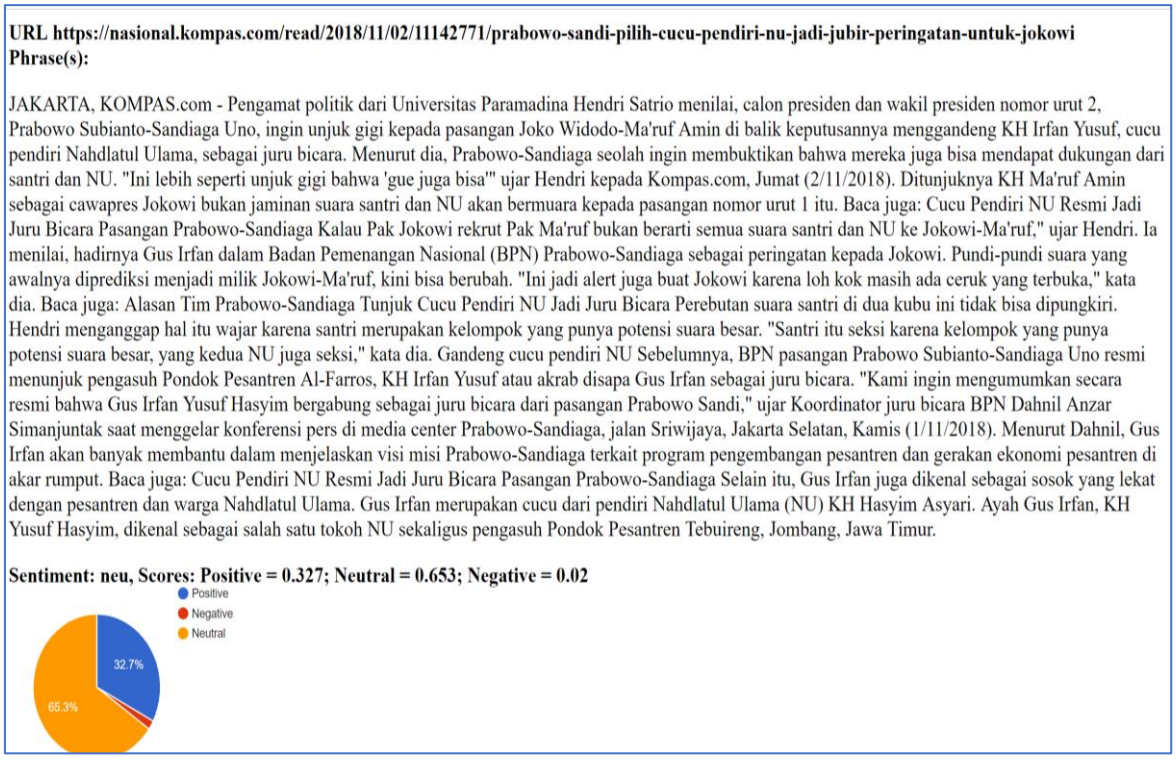
Sentiment: pos, Scores: Positive = 0.992; Neutral = 0; Negative = 0.008

Gambar 7. Contoh teks lengkap berita Jokowi-Maruf ID URL 3

Teks berita dengan ID URL 3 pada Gambar 7 di atas bertopik "jokowi maruf", dengan judul berita "*Jokowi-Ma'ruf Targetkan Kemenangan 60 Persen di Jawa Barat*", tersusun atas 292 kata. Berdasarkan evaluasi Naïve Bayes berita ini dianggap mengandung sentimen 0.992 (99,2%) positif, 0 (0%) netral, dan 0.008 (0,8%) negatif. Pemilihan kata-kata positif seperti "menang", "sinergi", "unggul"

ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

dan sejenisnya yang dilakukan penulis berita ini mendominasi atas kata-kata bermakna netral atau bahkan negatif. Akibatnya, sentiment berita ini memiliki kecenderungan positif yang kuat.



URL https://nasional.kompas.com/read/2018/11/02/11142771/prabowo-sandi-pilih-cucu-pendiri-nu-jadi-jubir-peringatan-untuk-jokowi
Phrase(s):

JAKARTA, KOMPAS.com - Pengamat politik dari Universitas Paramadina Hendri Satrio menilai, calon presiden dan wakil presiden nomor urut 2, Prabowo Subianto-Sandiaga Uno, ingin unjuk gigi kepada pasangan Joko Widodo-Ma'ruf Amin di balik keputusannya menggandeng KH Irfan Yusuf, cucu pendiri Nahdlatul Ulama, sebagai juru bicara. Menurut dia, Prabowo-Sandiaga seolah ingin membuktikan bahwa mereka juga bisa mendapat dukungan dari santri dan NU. "Ini lebih seperti unjuk gigi bahwa 'gue juga bisa'" ujar Hendri kepada Kompas.com, Jumat (2/11/2018). Ditunjuknya KH Ma'ruf Amin sebagai cawapres Jokowi bukan jaminan suara santri dan NU akan bermuara kepada pasangan nomor urut 1 itu. Baca juga: Cucu Pendiri NU Resmi Jadi Juru Bicara Pasangan Prabowo-Sandiaga Kalau Pak Jokowi rekrut Pak Ma'ruf bukan berarti semua suara santri dan NU ke Jokowi-Ma'ruf," ujar Hendri. Ia menilai, hadirnya Gus Irfan dalam Badan Pemenangan Nasional (BPN) Prabowo-Sandiaga sebagai peringatan kepada Jokowi. Pundi-pundi suara yang awalnya diprediksi menjadi milik Jokowi-Ma'ruf, kini bisa berubah. "Ini jadi alert juga buat Jokowi karena loh kok masih ada ceruk yang terbuka," kata dia. Baca juga: Alasan Tim Prabowo-Sandiaga Tunjuk Cucu Pendiri NU Jadi Juru Bicara Perebutan suara santri di dua kubu ini tidak bisa dipungkiri. Hendri menganggap hal itu wajar karena santri merupakan kelompok yang punya potensi suara besar. "Santri itu seksi karena kelompok yang punya potensi suara besar, yang kedua NU juga seksi," kata dia. Gandeng cucu pendiri NU Sebelumnya, BPN pasangan Prabowo Subianto-Sandiaga Uno resmi menunjuk pengasuh Pondok Pesantren Al-Farros, KH Irfan Yusuf atau akrab disapa Gus Irfan sebagai juru bicara. "Kami ingin mengumumkan secara resmi bahwa Gus Irfan Yusuf Hasyim bergabung sebagai juru bicara dari pasangan Prabowo Sandi," ujar Koordinator juru bicara BPN Dahnil Anzar Simanjuntak saat menggelar konferensi pers di media center Prabowo-Sandiaga, jalan Sriwijaya, Jakarta Selatan, Kamis (1/11/2018). Menurut Dahnil, Gus Irfan akan banyak membantu dalam menjelaskan visi misi Prabowo-Sandiaga terkait program pengembangan pesantren dan gerakan ekonomi pesantren di akar rumput. Baca juga: Cucu Pendiri NU Resmi Jadi Juru Bicara Pasangan Prabowo-Sandiaga Selain itu, Gus Irfan juga dikenal sebagai sosok yang lekat dengan pesantren dan warga Nahdlatul Ulama. Gus Irfan merupakan cucu dari pendiri Nahdlatul Ulama (NU) KH Hasyim Asyari. Ayah Gus Irfan, KH Yusuf Hasyim, dikenal sebagai salah satu tokoh NU sekaligus pengasuh Pondok Pesantren Tebuireng, Jombang, Jawa Timur.

Sentiment: neu, Scores: Positive = 0.327; Neutral = 0.653; Negative = 0.02

Gambar 8. Contoh teks lengkap berita Prabowo-Sandi ID URL 4

Gambar 8, ID URL 4 adalah berita bersumber dari keyword "prabowo sandi", berjudul *"Prabowo-Sandi Pilih Cucu Pendiri NU Jadi Jubir, Peringatan untuk Jokowi",* tersusun atas 368 kata. Berita ini memiliki skor sentimen 0,327 (32,7%) positif, 0,02 (2%) negatif, dan 0,653 (65,3%) netral. Dibanding berita sebelumnya, pada berita ini penulis tidak banyak memakai kata-kata bermakna positif ataupun negatif. Akibatnya kata-kata netral yang sejatinya berupa kata-kata umum yang mengonstruksi kalimat demi kalimat menjadi lebih dominan. Hasilnya, berita ini dianggap netral. Data lengkap hasil analisis tiap berita untuk masing-masing keyword ditunjukkan pada Tabel 5.

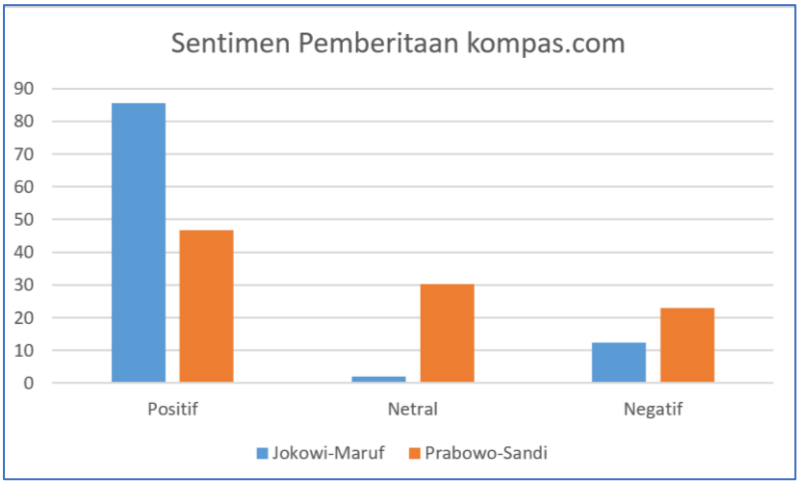
Tabel 5. Rekap 17 sentimen berita Capres-Cawapres 2019 pada Kompas.com.
(Data diambil per 13 Nopember 2018)

Skor sentimen "jokowi maruf"				Skor sentimen "prabowo sandi"			
ID URL	Positif	Netral	Negatif	ID URL	Positif	Netral	Negatif
1	1	0	0	1	0.029	0.057	0.914
2	0,571	0,143	0,286	2	0.842	0.105	0.053
3	0,992	0	0,008	3	0.327	0.653	0.02
4	1	0	0	4	0.32	0.64	0.04
5	0,994	0,002	0,004	5	0.571	0.286	0.143
6	1	0	0	6	0.108	0.865	0.027
7	0,955	0,015	0,03	7	0.842	0.105	0.053
8	0,33	0,01	0,66	8	0.977	0.008	0.015
				9	0.199	0.006	0.795
Total skor	6,841	0,17	0,988	Total skor	4.215	2.725	2.06
Rata-rata	0,855	0,021	0,124	Rata-rata	0.468	0.303	0.229

Fakta lain yang bisa diungkap dari percobaan ini adalah bahwa kedua keyword telah diberitakan secara positif oleh kompas.com. Namun skor rata-rata masing-masing keyword berbeda. Berita "jokowi maruf" berkecenderungan positif sebesar 0.855 atau 85%. Sedangkan kecenderungan sentiment positif

ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

untuk "prabowo sandi" adalah 0,468 atau 46,8%, sisanya cenderung netral atau negatif, seperti diperlihatkan pada Gambar 9.



Gambar 9. Sentimen berita Capres-Cawapres 2019 Kompas.com
Per 13 November 2018

4. Kesimpulan

Dari pengujian yang dilakukan dapat disimpulkan bahwa *topic-driven crawler* bekerja efektif dan efisien karena berhasil mengambil data web sesuai topik yang dibutuhkan saja. Begitu pula filter *text preprocessing*, bagian ini telah mampu memisahkan teks target dari kode atau karakter yang tidak diperlukan. Hasilnya, teks berita dapat dianalisis dengan hasil yang cukup jelas. Dari 17 berita di kompas.com yang dianalisis, kedua pasangan Capres-Cawapres telah diberitakan positif. Namun, Jokowi-Ma'ruf cenderung diberitakan lebih positif daripada Prabowo-Sandi.

Sebagai saran penyempurnaan, untuk mendapatkan berita terbaru, query Google yang ditanamkan pada crawler dapat custom berdasarkan indeks terbaru harian, mingguan, bulanan, atau tahunan. Sehingga data yang diambil selain yang relevan juga terbaru. Agar proses pengambilan data dapat berjalan otomatis, crawler dapat dieksekusi secara terjadwal pada level sistem operasi, misalnya menggunakan cron pada sistem operasi Linux.

Referensi

[1] Wira Respati, "Transformasi Media Massa Menuju Era Masyarakat Informasi di Indonesia", Humaniora Vol. 5, 2014.
[2] Thao Ton, "Political Marketing in the Digital Era: Millennials' use of Social Media for Political Information and Its Effect on Voting Decision", DePaul University 2016.
[3] Bing Liu, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, 2012.
[4] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, *From Data Mining to Knowledge Discovery in Databases*, AI Magazine, Vol. 17, No. 3, 1996.
[5] Sholom M. Weiss, Cs., *Text Mining: Predictive Methods for Analyzing Unstructured Information*, Springer, 2004.
[6] S. Vijiyarani, Ms E. Suganya, "Research Issues in Web Mining", International Journal of Computer-Aided Technologies (IJCAx), Vol. 2, No.3, 2015.
[7] Preeti Rathi, Nipur Singh, "A Survey of Issues and Techniques of Web Usage Mining", International Research Journal of Engineering and Technology, Vol. 04, No. 07, 2017.
[8] Junghoo Cho, Hector Garcia-Molina, "Parallel Crawlers", ACM 1-58113-449-5/02/0005.
[9] Filippo Menczer, Cs, "Evaluating Topic-Driven Web Crawlers", SIGIR'01, 2001.
[10] Bing Liu, "Opinion Observer: Analyzing and Comparing Opinions on the Web", ACM 1-59593-046-9/05/0005.
[11] Pranali Yenkar, S.D. Sawarkar, "Opinion Mining of the Movie Blogs based on Supervised Learning Approach", International Journal of Advanced Research in Computer Science, Volume 4, No. 1, 2013.
[12] Seyed M. Mirtaheri, Cs., "A Brief History of Web Crawlers", IBM Canada Ltd., 2013

ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

[13] Vivek Narayanan, Cs., "Fast and accurate sentiment classification using an enhanced Naive Bayes model", Indian Institute of Technology, 2013.
[14] Bryan Nii Lartey, Cs., "Web Application for Sentiment Analysis Using Supervised Machine Learning", International Journal of Software Engineering and Its Applications Vol. 9, No. 1, 2015.
[15] Guangren Shi, *Data Mining and Knowledge Discovery for Geoscientists*, Elsevier, 2014.
[16] Wahid, D. H., & Azhari, S. N., "Peringkasan Sentimen Esktraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity". IJCCS, Vol. 10, No. 2, 207-218, 2016.